自适性语言测试题库 CTT 与 IRT 参数估计对比研究

修美丽 柴省三 鲍 杰

摘 要: 在计算机自适应性(CAT)与评价中,题库是测试系统实现的基础,其中题目参数估计质量是决定题库建设的核心和必要条件。研究以 22582 名中国汉语水平考试(HSK)考生的实测大数据为基础,采用分层随机抽样法,对获得的 8 个不同容量的被试样本进行了 CTT 和 IRT 参数的估计及比较研究。结果发现,在样本分布同质性前提下,当样本规模在 500 以内时,CTT 与 IRT 的项目难度参数、区分度参数之间一致性较差;项目的区分度参数稳定性不高;考生能力参数之间存在显著相关性。在远程自适应性语言测试的题库建设中,建议采用大样本 IRT 参数估计方案。

关键词: 计算机自适应性考试;题库参数估计;教育测量;语言测试;CTT 与 IRT 理论
中图分类号:H09 **文献标识码:**A **文章编号:**1004-9142(2024)03-0180-10

一、问题的提出

随着计算机科学和人工智能(AI)技术的普及与发展,以经典测量理论(Classical Test Theory,简称 CTT)为基础的网络考试和以项目反应理论(Item Response Theory, 简称 IRT)为基础的计算机自适应性考试(Computer Adaptive Testing, 简称 CAT)已经成为智能教育测量和语言测试的重要手段之一。基于大数据对教育现象的关键要素及其相互关系进行定量刻画和价值判断的方式正广泛应用于教育研究、决策与实践中[①②]。为了推广基于国际互联网的远程教育测量模式,为全球考生提供便捷的考试方式,

作者简介:修美丽,对外经济贸易大学国际学院讲师;
柴省三,北京语言大学国际学生教育政策与评价研究院教授;
鲍杰,北京语言大学国际学生教育政策与评价研究院博士研究生。

基金项目:国家社科基金重点项目(17AYY011)、北京语言大学重大基础研究专项(中央高校基本科研业务费专项资金)(19ZDJ04)、北京语言大学一流学科团队支持计划(2023YJC02)阶段性成果。

① 范涌峰、宋乃庆:《大数据时代的教育测评模型及其范式构建》,《中国社会科学》2019 第 12 期。

② 魏非、宫玲玲、章玉霞、祝智庭:《基于微能力的教师信息化教学能力测评模型》,《现代远程教育研究》2022 年第 6 期。

美国教育测量服务中心(ETS)开发的英语作为外语的语言水平测试(TOEFL)已经于 2005 年正式推出了网络版 TOEFL 测量模式(TOEFL iBT)①。中国汉语水平考试(Hanyu Shuiping Kaoshi，简称 HSK)在 1999 年就开始探索题库建设的可行性问题②,题库建设的参数采集标准与描述问题③,以及完善 HSK 题库参数的层次、体系及价值问题④。全国大学英语四、六级(CET)考试委员会也从 2008 年开始进行了若干次远程网络考试的实验研究,为全面实现 CET 考试的远程化、网络化和自适应性考试模式奠定了基础。

无论是在计算机辅助远程测试还是计算机自适应性智能测试中,参数标定准确、稳定的题库(item bank)是网络考试系统题目自动选择实现的核心要求,也是影响网络测试信度(reliability)和效度(validity)高低的重要因素⑤⑥⑦。CTT 和 IRT 是现代教育测量和语言测试领域中题目参数估计和考生能力评估的两大主要理论体系⑧⑨⑩。其中,CTT 模型相对简单、计算方便,在题目参数估计、题库建设、分数处理等方面已经较为成熟和完善。IRT 则是一种较新的,完全不同于 CTT 的教育测量模型,它在题目参数估计、考生能力参数估计和题库建设,以及计算机自适应考试等方面具有其自身的优势。尽管 CTT 和 IRT 在国内外大规模、标准化语言测试中均得到了广泛应用⑪⑫⑬⑭,但是两者在参数估计的样本依赖性、参数一致性方面仍存在较大争议⑮⑯。

学界从不同角度尝试对 CTT 和 IRT 进行了的比较研究⑰⑱,但是由于不同的研究者所采用的测验类型、研究材料、样本规模、IRT 模型以及参数估计方法存在差异,研究结论也不完全相同。有的学者认为 IRT 在多方面显著优于 CTT⑲;而有些学者则认为 IRT 的参数不变性在实证研究中不能满足⑳㉑。现有研究的研究设计或多或少存在研究数据不规范、研究样本控制不够严谨等问题㉒。与此同时,基于实测大规模第二语言测试的 CTT 和 IRT 参数估计的实证研究也比较缺乏。

---

① Fulcher, G. , *Practical Language Testing*, London: Hodder Education, 2010.

② 谢小庆、任杰:《对从 HSK 题库中计算机自动生成试卷稳定性的试验检验》,《心理学探新》1999 年第 4 期。

③ 张凯:《汉语水平考试题库的描述性参数》,《世界汉语教学》1999 年第 3 期。

④ 杨翼:《HSK(高等)题库参数体系研究》,《语言教学与研究》2006 年第 4 期。

⑤ 蔡旻君、刘仁云:《计算机辅助教育测量与评价》,中国水电出版社 2010 年版。

⑥ 杨建原、柏桧、赵守盈:《计算机自适应性测验开发的程序研究》,《中国考试》2012 年第 3 期。

⑦ Chalhoub-Deville M. and Deville C. ,“ Computer Adaptive Testing in Second Language Contexts, ”*Annual Review of Applied Linguistics*, vol. 19, 1999, pp. 273-299.

⑧ 罗照盛:《项目反应理论基础》,北京师范大学出版社 2012 年版。

⑨ Hambleton R. K. and Jones R. W. ,“Comparison of Classical Test Theory and Ttem Response Rheory and their Applications to Test Development;” *Educational Measurement: Issues and Practice*, vol. 12, no. 3, 1993, pp. 38-47.

⑩ 钱锦昕:《基于项目反应理论和量子智能算法的选题策略研究》,东南大学出版社 2022 年版。

⑪ 漆书青、戴海崎、丁树良:《现代教育与心理测量学原理》,高等教育出版社 2002 年版。

⑫ 杜文久:《心理与教育统计测量》,西南师范大学出版社 2009 年版。

⑬ 黄光扬:《教育测量与评价》,华东师范大学出版社 2003 年版。

⑭ 金娣、王钢:《教育评价与测量》,教育科学出版社 2007 年版。

⑮ 漆书青:《现代测量理论在考试中的应用》,华中师范大学出版社 2003 年版。

⑯ Jabrayilov R. , Emons, W. H. and Sijtsma K. ,“ Comparison of Classical Cest Cheory and Item Response Theory in Individual Change Assessment, ”*Applied Psychological Measurement*, vol. 40, no. 8, 2016, pp. 559-572.

⑰ Hernandez R. , “Comparison of the Item Discrimination and Item Difficulty of the Quick-Mental Aptitude Test Using CTT and IRT Methods,” *The International Journal of Educational and Psychological Assessment*, vol. 1, no. 1, 2009, pp. 12-18.

⑱ 张军:《单维参数型与非参数型项目反应理论项目参数的比较研究》,《心理学探新》2015 年第 3 期。

⑲ 陈谨、何静、佟仁城:《英语标准化考试评价中 IRT 与 CTT 的比较研究》,《数学的实践与认识》2011 年第 20 期。

⑳ Cook L. L. Eignor, D. R. and Taft H. L. ,“A Comparative Study of the Effects of Recency of Instruction on the Stability of IRT and Conventional Item Parameter Estimates, ”*Journal of Educational Measurement*, vol. 25, no. 1, 1988, pp. 31-45.

㉑ Fan X. ,“Item Response Theory and Classical TestTheory: An empirical Comparison of their Item/Person Statistics,” *Educational and Psychological Measurement*, vol. 58, no. 3, 1998, pp. 357-381.

㉒ Garcia M. I. B. , Marañón, P. P. , Falcón, J. C. S. and Costas C. S. ,“Empirical Relationships between Classical Test Theory and Item Response Theory Statistics, ” *Psicothema*, vol. 13, no. 2, 2001, pp. 342-329.

中国汉语水平考试(HSK)是一种典型的大规模、标准化的常模参照性语言测试。它是由北京语言大学设计的以测试母语非汉语者(包括来华留学生、华侨和国内少数民族考生)的汉语水平为目的的国家级标准化考试。中国汉语水平考试(HSK)成绩是来华留学生进入中国高等院校接受本科、研究生学历教育所必须的语言水平证明,同时也是一些在华商务机构职员选拔、录用时汉语水平高低的重要评估依据①,自 1989 年开考以来已积累的大量测验题目为题库建设和智能测试系统的设计奠定了基础。有别于试卷库,题库既不是以整套试卷为单位的试卷集合,也不是若干试题或题目的简单积累②,而是经过实测后根据不同教育测量理论对具体题目或题库(item bundle)估计参数、标定测量目标等若干属性并经过参数等值(equating)后的标准化题目集合。题库质量与参数估计的准确性、稳定性和属性标定的完整性以及参数等值的误差控制直接相关③。题库是实现计算机自适应性考试自动选题的基础,是智能组卷的关键要素。高质量的题库对中国汉语水平考试(HSK)、高考招生英语考试④、高中学业水平考试⑤等高利害性(high stakes)教育测量的信度和效度具有重要的影响。中国汉语水平考试(HSK)是面向国际学生的大型标准化考试,如果在如此重要的国家级教育测验中存在试题分析问题和考生特质水平等参数估计问题,那么必然影响计算机自动组卷系统、计算机自适应性测试系统的题目选择质量,影响题库(item bank)参数标注的一致性以及平行试卷(parallel test)之间的难度平衡性,同时测量结果对考生来说无疑也是不公平的,最终考生对教育考试也会产生“信任危机”。

鉴于上述 CTT 和 IRT 在理论和实际应用中存在的诸多争议,以及对以中国汉语水平考试(HSK)为例的常模参照类语言测验的实证研究上存在的不足。本研究将以中国汉语水平考试(HSK)实测大数据(big data)为基础,在严格控制样本质量的前提下,以 8 个不同样本容量的中国汉语水平考试(HSK)听力测试数据为研究对象,专门对中国汉语水平考试(HSK)的 CTT 与 IRT 项目参数和能力参数之间的一致性问题及参数的样本依赖性问题进行实证研究,以期为大规模教育测量与评价的项目参数估计、题库建设、等值模型(equation model)和计算机自适应性考试系统的设计提供理论借鉴。

二、研究设计

为了克服以往研究中存在的诸多局限性,在对 CTT 和 IRT 参数进行估计和比较性研究时,研究设计最好满足以下 4 个条件,其一是使用大规模语言测试实测数据,其二是测试试卷的信度和效度满足较高的要求,其三是样本容量之间具有较大跨度,样本之间具有较高的同质性,其四是使用性能稳定的参数估计软件。因此,本研究的具体思路如下:首先,对中国汉语水平考试(HSK)的原始实测数据进行预处理,通过自编程序和统计软件获得 22582 名考生在听力分测验(50 个题目)上的作答数据,然后,采用分层随机抽样方法(Stratified random sampling)抽取 8 个不同样本容量的研究样本(200 人、300 人、500 人、800 人、1000 人、2000 人、5000 人、7000 人);其次,分别对这 8 个样本进行 CTT 和 IRT 分析,估算基于 CTT 的项目参数、能力参数以及基于 IRT 的项目参数、能力参数;最后,运用皮尔逊积差相关分析法对 CTT 和 IRT 参数之间的一致性进行比较研究,运用单因素方法分析法(one-way ANOVA)对 CTT 和 IRT 参数的跨样本不变性进行检验。

(一)研究被试

研究数据来自 2005 年 12 月参加 HSK 的 22582 名考生的实测数据。在这些考生中,男生共 10468

① 柴省三:《汉语水平考试(HSK)复本测验稳定程度的历时性研究》,《现代语文》2011 年第 2 期。

② 国家教委考试中心:《题库建设理论与实践》,光明日报出版社 1991 年版。

③ 杨志明:《打通题库建设瓶颈的关键手段——基 Rasch 模型的题目参数等值技术》,《教育测量与评价》2017 年第 1 期。

④ 徐良:高考英语(上海卷)题库命题背景下的校本题库建设——基于语法填空题的案例研究》,《外语测试与教学》2018 年第 1 期。

⑤ 马鹏飞、李建军:《新高考背景下学业水平考试题库建设的现状、功能及特征》,《考试研究》2022 年第 3 期。

人,占比为 46.36%,女生共 12114 人,占比为 53.64%。他们分别来自亚洲、欧洲、美洲、大洋洲和非洲的 110 个国家,母语背景达 66 种。因此,考生的来源具有明显的多样性和代表性。

(二)研究材料

本研究采用中国汉语水平考试(HSK [初中等],卷标为 M0590X)的实测数据进行 CTT 和 IRT 的参数估计和比较研究。中国汉语水平考试(HSK)经过 20 多年的测试实践检验,具有比较理想的信度(reliability)和效度(validity)特征①②③。HSK[初中等]考试由听力理解(50 题)、语法(30 题)、阅读理解(50 题)和综合填空(40 题)四个分测验组成。所有试题均为客观题,采用 0/1 计分的方法,原始分满分为 170 分。考虑到 IRT 在参数估计时对测验的单维性(unidimensional)假设方面具有较高的要求④⑤⑥,因此,本研究仅针对 HSK 的听力测验部分进行实证研究,以确保参数估计的准确性。分测验及全卷统计指标与信度系数(Cronbach α 系数)估计值见表 1。

表 1　各分测验的基本描述

	题数	均值	标准差	偏态	峰度	信度系数
听力	50	30.16	10.15	-0.228	-0.916	0.911
语法	30	17.41	6.60	-0.044	-0.970	0.868
阅读	50	28.51	10.97	-0.095	-1.029	0.924
综合填空	40	22.06	8.39	-0.166	-0.833	0.892
全卷	170	98.15	33.44	-0.115	-0.985	0.971

由表 1 可知,此次测验无论是从分测验来看还是全卷来看,信度系数都比较高。其中,全卷信度系数高达 0.971。各个分测验的信度系数分别为:0.911(听力分测验),0.868(语法分测验),0.924(阅读分测验),0.892(综合填空分测验)。以上结果说明本次测验具有较高的稳定性、一致性和可靠性。

(三)研究工具

本研究的主要材料是 HSK 听力测验的作答数据。我们首先通过 Foxpro8.0 将考生的作答结果进行整合,通过自编程序将考生直接作答结果资料转化为 0/1 计分数据,同时计算各个分测验的总分。然后,运用 SPSS18.0 进行分层随机抽样及 CTT 的参数估计。最后,运用西南师范大学心理学系开发的 IRTP 软件进行 IRT 参数估计⑦⑧。参数估计采用的是三参数 IRT 逻辑斯蒂(logistic)模型。被试在某测验项目上的答对概率 $P(\theta)$ 与该测验项目的区分度参数 a、难度参数 b 以及伪机遇水平参数(Pseudo-chance Level Parameter),即猜测参数 c 之间的关系见公式 1。

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + \exp[1.7a_i(\theta - b_i)]} \qquad \text{(公式 1)}$$

① 柴省三:《汉语水平口试信度的理论与实证研究》,《语言教学与研究》2003 年第 4 期。

② 柴省三:《汉语水平考试(HSK)复本测验稳定程度的历时性研究》,《现代语文》2011 年第 2 期。

③ 谢小庆:《对 15 种测验等值方法的比较研究》,《心理学报》2000 年第 2 期。

④ Hambleton R. K. and Jones, R. W. ,"Comparison of Classical Test Theory and Item Response Theory and their Applications to Test Development," *Educational Measurement: Issues and Practice*, vol. 12, no. 3,1993, pp. 38-47.

⑤ Szabó G., *Applying Item ResponseTheory in Language Test Item Bank Building*, Bern, Switzerland: Peter Lang. 2008.

⑥ 丁树良、罗芬、涂冬波:《项目反应理论新进展专题研究》,北京师范大学出版社 2012 年版。

⑦ 杜文久:《心理与教育统计测量》,西南师范大学出版 2009 年版。

⑧ 杜文久:《高等项目反应理论》,科学出版社 2016 年版。

(四)研究方法

在估算出 CTT 和 IRT 的项目参数和能力参数后,我们采用皮尔逊积差相关法对 CTT 的原始分数与 IRT 的考生能力参数 θ,CTT 的项目参数与 IRT 的项目参数进行相关分析,据此考察在不同样本条件下估算出的 CTT 参数与 IRT 参数的一致性程度。然后,采用单因素方差分析法对基于 CTT 理论和 IRT 理论的 8 个样本估算出的项目参数、能力参数进行差异性检验,考察 CTT 参数、IRT 参数的跨样本一致性(parameters invariance across samples)程度,据此研究 CTT 和 IRT 的参数不变性。

三、数据预处理与分析

数据预处理阶段包括选择抽样方案、样本方差齐性检验、样本分布正态性检验、模型拟合度检验以及难度参数转换五个步骤。

(一)抽样方案

通过对考生 HSK 听力理解测验分数的统计发现,处于[0,9]这个分数段的考生非常少,仅占考生总体的 1.13%左右,而且此分数区间内的考生得分多为 8 分和 9 分。考虑到原始分数分布的特点,我们只考察了处于[10,50]这个分数段的考生听力测验数据。

本研究采用分层随机抽样方法抽取研究样本。我们首先将听力测验分数分为四个层次,分别是 10-20 分、21-30 分、31-40 分、41-50 分。然后按照正态分布的近似比例,分别从这四个分数段中随机抽取大约 15%、35%、35%和 15%的考生听力测试数据。为了确保分数均衡分布且接近于正态分布①②,在每个分数段中对个别分数的频数分布进行了多次调整。最终将各个样本中的 4 个分层随机抽样结果组合在一起,由此获得我们要进行参数估计与比较的 8 个不同规模的研究样本。

(二)样本方差齐性检验

为了更好地比较 CTT 和 IRT 两种参数之间的一致性及与样本容量之间的关系,我们必须严格控制样本的质量,确保不同样本之间满足方差齐性之要求。

方差齐性检验(Homogeneity of variance test)是数理统计中考查不同样本的总体方差是否相同的一种方法。如果不同样本通过检验后证明方差齐,那么说明这些样本在很大程度上可能来自同一总体;反之,如果通过检验后证明方差不齐,那么在此样本基础上进行的研究将可能得出错误的结论。常见的方差齐性检验方法有:Bartlett 检验法、Hartley 检验法、Brown-Forsythe 检验法、Welch 检验法和 Levene 检验法等③。本研究采用的方差齐性检验方法是 Levene 检验法。针对 8 个样本进行的方差齐性检验结果和方差分析结果分别见表 2 和下页表 3。

表 2　方差齐性检验

Levene 值	自由度 1	自由度 2	*P* 值
1.349	7	16792	0.223

由表 2 的方差齐性检验结果可知,*P* 值为 0.223(>0.01),明显大于 0.01,我们应该接受零假设,拒绝备择假设(H1),即这 8 个样本是来自同一总体,样本的同质性较高,样本满足方差齐性的基本要求。

① 吴明隆、涂金堂:《SPSS 与统计应用分析》,东北财经大学出版社 2012 年版。
② 张厚粲、徐建平:《现代心理与教育统计学》,北京师范大学出版社 2020 年版。
③ 吴明隆、涂金堂:《SPSS 与统计应用分析》,东北财经大学出版社 2012 年版。

表 3　方差分析结果

	平方和	自由度	均方	*F* 值	*P* 值
组间差异	80. 192	7	11. 456	0. 154	0. 994
组内差异	1252190. 091	16792	74. 571		
合计	1252270. 283	16799			

表 3 的方差分析结果显示，*F* 值未达到显著性水平（*P* = 0. 994），因此接受零假设，可以认为 8 个样本的平均数之间不存在显著性差异。

（三）样本正态性检验

为避免因样本分数分布存在较大差异而对参数估计产生系统性影响，在上述方差齐性检验的基础上，本研究拟选择测验分数符合近似正态分布的样本进行参数估计。通过 Kolmogorov-Smirnov 分布检验，8 个样本的频数分布与正态曲线具有高度的拟合性，所以研究样本满足正态分布的基本要求。

（四）模型拟合检验

本研究采用逻辑斯蒂模型对数据进行拟合，采用博克的 χ^2 检验法进行拟合性检验①②③。经检验发现本次 HSK 听力分测验的 50 道试题中总共有 7 道试题（第 3、6、7、14、19、20、42 题）与 IRT 模型拟合较差，即实际卡方值大于临界值，因此要拒绝零假设 H_0（逻辑斯蒂模型与数据集是拟合的），从而接受备择假设 H_1（逻辑斯蒂模型与数据集不拟合）。通过拟合曲线可以发现，只有第 19 题受到极端值的影响，其卡方值较大，拟合较差，其余试题除了尾部拟合略差以外，其他部分均拟合良好。考虑到考生总量非常大，容易对卡方值造成不良影响，同时，50 道听力试题均为选择题，被试在作答过程中难免会有猜测成分。因此我们有理由认为该数据和三参数逻辑斯蒂模型是拟合的，可以进行参数不变性对比研究。

（五）难度参数转换

CTT 的难度参数表示在 0/1 计分中全体考生在这个题目上的答对率，其数值介于 0 和 1 之间，数值越小，表示答对该题的人越少，题目越难；数值越大，则表示答对该题的人越多，测验题目对该组被试而言就越简单④⑤。CTT 的难度值实际上是“易度”（facility）。

IRT 的难度参数表示项目特征曲线（Item Characteristic Curve，简称 ICC）的拐点在能力特质水平轴上的取值。因此 IRT 的难度代表的是“真正的难度”，即 IRT 的难度参数（*b*）越大，题目越难；IRT 的难度参数越小，题目越简单。所以，CTT 和 IRT 的难度参数在数值上代表的是相反的概念。鉴于此，为了进行 CTT 和 IRT 的难度参数一致性研究，我们采用了转换 CTT 难度的方法，也就是用 1 减去 CTT 的难度参数值得出一个与 IRT 难度属性相同的值，以便将该值与 IRT 的难度参数进行相关分析和方差分析。为行文方便，本文所指的 CTT 难度参数统一定义为转换后的难度值。

---

① Garcia M. I. B., Marañón, P. P., Falcón, J. C. S. and Costas C. S. c., “Empirical Relationships between Classical Test Theory and Item Response Theory Statistics,” *Psicothema*vol. 13, no. 2, 2001, pp. 342-329.

② Hambleton, R. K., and Jones, R. W., “Comparison of Classical Test Theory and Item Response Theory and their Applications to Test Development,” *Educational Measurement: Issues and Practice*, vol. 12, no. 3, 1993, pp. 38-47.

③ Szabó, G., *Applying Item Response Theory in Language Test Item Bank Building*, Bern, Switzerland: Peter Lang. 2008.

④ 漆书青、戴海崎、丁树良：《现代教育与心理测量学原理》，高等教育出版社 2002 年版。

⑤ Bachman, L. F., *Fundamental Considerations in Language Testing*, Oxford: Oxford University Press. 1990.

四、研究结果与分析

基于相同样本所估计的 CTT 参数与 IRT 参数之间是否一致,以及 CTT 参数、IRT 参数分别在不同样本之间是否具有跨样本一致性问题一直以来都是学界争论的话题①②。因此,本研究首先针对 8 个不同样本在 CTT 参数和 IRT 参数之间的一致性问题进行了分析,然后对 CTT 参数和 IRT 参数的跨样本一致性及参数不变性分别进行了研究。

(一)项目参数和能力参数一致性研究

为了考察 CTT 和 IRT 在 8 个样本中估计出的项目参数和考生能力参数之间的一致性程度。我们分别将 CTT 的难度参数与 IRT 的难度参数(见表 4)、CTT 的区分度参数与 IRT 的区分度参数(见表 5)、CTT 的原始分数与 IRT 的能力参数 θ(见表 6)进行了皮尔逊积差相关分析。

表 4　CTT 难度参数和 IRT 难度参数相关分析结果③

样本容量	CTT 和 IRT 的难度参数相关
200	0.344*
300	0.503**
500	0.415**
800	0.832**
1000	0.927**
2000	0.887**
5000	0.925**
7000	0.830**

通过表 4 可知,从整体上来看,基于这 8 个不同规模的样本所估计出的 CTT 和 IRT 的难度参数之间均在 0.05 水平上存在显著的正相关关系。尤其是当样本规模较大时($n \geqslant 800$),两者的难度参数之间的相关系数均在 0.800 以上。其中,当样本规模为 1000 和 5000 时,两者的难度相关系数达到了 0.900 以上。但是当样本规模较小时,二者虽然也存在显著正相关,但是相关系数较低,比如当样本容量为 200 时,二者的相关系数仅为 0.344。这说明,当在考生人数较多的大规模标准化语言测试中,利用 CTT 和 IRT 所估计出来的难度参数相关性较高,难度参数之间的一致性较强。

表 5　CTT 和 IRT 区分度参数的相关分析结果

样本规模	CTT 和 IRT 区分度参数相关系数
200	0.342*
300	−0.022(.490**)
500	0.065(.550**)

① Szabó, G., *Applying Item Response Theory in Language Test Item Bank Building*, *Bern*, Switzerland: Peter Lang. 2008.

② De La Torre, J., and Deng, W., *Improving person-fit Assessment by Correcting the Ability Estimate and its Reference Distribution*, Journal of Educational Measurement, vol. 45, no. 2, 2008, pp. 159−177.

③ 本文表 4、表 5 和表 6 中的“*”表示相关在 0.05 水平上显著,“**”表示相关在 0.01 水平上显著。

续表5

样本规模	CTT 和 IRT 区分度参数相关系数
800	0.497**
1000	0.374**
2000	0.448**
5000	0.361**
7000	0.468**

由表 5 可知,总体上来看,除了样本规模为 300 和 500 的研究样本外,其余 6 个不同容量的样本所估计出的 CTT 和 IRT 的区分度参数之间都呈显著正相关,但是相关系数都不高,均在 0.500 以下。

当样本规模为 300 和 500 时,CTT 和 IRT 的区分度参数相关性并不理想,相关系数分别为-0.022 和 0.065。也就是说,一个是负相关,一个相关近似于 0。这两个数值与其他数值存在明显不同。究其原因,当我们回到原始参数估计值时就会发现,在样本容量为 300 时,第 22 题的 IRT 的区分度参数估计值 a 为 61.14,属于极端异常值。出现这一结果的原因可能与本研究所采用的 IRTP 参数估计程序对区分度参数的估计值所设定的范围是(-∞ , +∞),及测验题目的质量有关。所以,鉴于此题的特殊性,我们删去第 22 题后,重新估计并计算了 CTT 和 IRT 区分度参数及两者之间的积差相关系数,最后所得结果为 0.490,此结果显示两者在 0.01 水平上显著相关。当样本规模为 500 时,相关系数出现异常的情况与样本规模为 300 的样本类似。不过,此时是第 6 题的 IRT 的区分度参数估计值出现了异常,其估计值为 18.26。我们采取同样的处理方法重新估计并计算了 CTT 和 IRT 的区分度参数和相关系数,所得结果为 0.550。此时 CTT 和 IRT 的区分度参数在 0.01 水平上呈显著正相关。

表 6　CTT 和 IRT 能力参数的相关分析结果

样本规模	CTT 和 IRT 的能力参数相关
200	0.873**
300	0.866**
500	0.875**
800	0.887*
1000	0.855*
2000	0.878*
5000	0.884**
7000	0.881**

由表 6 可知,整体来看,基于这 8 个不同规模的样本所估计出的 CTT 和 IRT 的考生能力参数之间都存在显著的正相关关系。而且两者之间的相关系数都在 0.800 以上,同时相关系数的差异不大,具有明显的跨样本一致性。这说明无论是基于小样本还是大样本,IRT 和 CTT 所估计的考生能力参数之间都具有很高的一致性。

(二)项目参数和能力参数不变性研究

所谓项目参数、能力参数的不变性,是指 CTT、IRT 参数在不同样本之间的参数不变性(parameter invariance cross samples)。本研究首先运用 IRTP 软件针对 8 个不同样本进行 CTT 和 IRT 项目参数(难

度、区分度)和考生能力参数(原始分、θ 值)的估计,然后采用单因素方差分析对其进行差异性检验,具体分析结果见表 7。

表 7　CTT 和 IRT 方差分析结果

			平方和	自由度	均方
CTT	难度	组间差异	0.003	7	0.000
		组内差异	8.851	392	0.023
		合　　计	8.854	399	
	区分度	组间差异	0.010	7	0.001
		组内差异	4.108	392	0.010
		合　　计	4.118	399	
	考生能力	组间差异	80.192	7	11.456
		组内差异	1252190.091	16792	74.571
		合　　计	1252270.283	16799	
IRT	难度	组间差异	17.875	7	2.554
		组内差异	1514.969	392	3.865
		合　　计	1532.845	399	
	区分度	组间差异	65.296	7	9.328
		组内差异	4002.597	392	10.211
		合　　计	4067.893	399	
	考生能力	组间差异	0.446	7	0.064
		组内差异	21303.192	16792	1.269
		合　　计	21303.639	16799	

由表 7 的方差分析结果可知,8 个不同规模的考生样本在 CTT 与 IRT 的项目参数(难度、区分度)和考生能力参数(原始分、θ 值)上均不存在显著差异($P>0.05$)。由此可见,CTT 与 IRT 的项目参数和考生能力参数在不同样本之间具有较好的参数不变性。

五、研究结论与建议

本实证研究以中国汉语水平考试(HSK)的实测大数据为基础,采用分层随机抽样方法,在严格控制样本分布的条件下,分别针对 CTT 理论和 IRT 理论估计出的项目参数(难度和区分度)和能力参数的一致性和跨样本不变性进行了对比研究。在考生样本数据基本满足近似正态性分布的前提下,采用 CTT 和 IRT 所估计的项目参数、考生能力参数之间具有不同性质、不同程度的一致性和参数不变性。研究结论分述如下。

第一,从难度参数来看,在 8 个不同容量样本上估计的 CTT 难度参数和 IRT 难度参数之间均存在显著正相关。当样本规模较小时,两者之间的相关程度相对较弱;当样本规模较大时,两者的相关程度较强;CTT 和 IRT 难度参数之间并没有表现出样本规模越大,相关程度就越高的稳定性趋势。因此我们建议:在大规模标准化远程语言测试题库建设的参数估计中,尽量采用样本容量在 1000 以上的被试团体估计难度参数,而且使用 CTT 与 IRT 估计的项目难度参数之间不存在显著差异。当样本容量小于

800 时,CTT 和 IRT 之间的参数一致性较差,究竟选择何种估计方法应慎重考虑。

第二,从区分度参数来看,在 8 个不同容量样本中,CTT 和 IRT 区分度参数之间均存在显著正相关。但是与上述难度参数之间的相关性相比,相关系数明显偏低。无论样本规模的大小,通过 CTT 和 IRT 估计的区分度参数之间的相关性都相差不大。但是,在运用 IRT 理论估计区分度参数时,IRT 区分度参数易受个别特殊(太难或太易)质量题目的影响;与 IRT 相比,CTT 估计的结果相对稳定。因此,在远程计算机自适应性语言测试的题库参数估计中,究竟采取 CTT 模型还是采用 IRT 模型,区分度参数是优先考虑的因素。如果采用 IRT 估计区分度参数,那么建议首先通过“预估计”方法在发现并排除异常质量题目后再估计题目参数,从而确保题库区分度参数的稳定性。

第三,从能力参数角度来看,通过对 8 个不同容量样本的研究,我们发现:CTT 的考生能力参数和 IRT 的考生能力参数之间均存在显著正相关性,而且二者之间的相关程度在 8 个样本之间具有跨样本一致性。可见,当测验分数满足近似正态分布要求时,无论 CTT 模型还是 IRT 模型都能较好地估计出考生的能力水平。不过,与 IRT 的能力值相比,CTT 的能力参数(分数)具有相对性,因此建议在采用 CTT 参数构建计算机自适应性考试题库时,应适当增加测验题目的数量,以保证测验结果具有较高的信度。

第四:从难度参数、区分度参数和能力参数总体来看,用 CTT 和 IRT 所估计出的指标之间的一致性程度存在较大差异。其中,CTT 能力参数与 IRT 能力参数之间的相关性最高,一致性最好,同时其一致性程度受样本规模的影响也最小。难度参数之间的相关性适中,但受样本规模大小的影响比较明显。区分度参数之间的相关性,不管样本规模如何,参数一致性最差,因此区分度参数是题库建设中最值得重视的指标。

题库质量是影响远程教育测量、计算机自适应性语言测试信度和效度的核心,评价题库质量的主要标准是题目的难度参数、区分度参数是否稳定。在题目参数估计的 CTT 模型和 IRT 模型选择中,学术界和语言测试机构从业者长期存在争议。本研究在 8 个不同样本上对 CTT 和 IRT 题目参数之间、能力参数之间的一致性进行了实证研究,同时分别针对 CTT 和 IRT 估计的题目(难度、区分度)参数、能力参数是否具有跨样本一致性进行了考察。

(责任编辑:子聿)